Tunisian republic
Ministry of Higher Education
and Scientific Research

University of Sfax

National School of Electronics% and
telecommunications of Sfax

Engineer in :
Industrial Computer Engineering

Report :
SUMMER INTERNSHIP

Year : 2023-2022

SUMMER INTERNSHIP REPORT

Introduced to

National school of Electronics and Telecommunications of Sfax

HOSTING COMPANY

DATACAMP TRAINING AND CONSULTING

Author

MANAI MOHAMED MORTADHA

Revolutionizing Vocal Track Extraction : Innovative Hybrid Neural Network Approaches with Deep Clustering, U-net, and UH-net Models

Mr. Ben Hamed Bassem — **Academic Supervisor**
Mr. Bouhamed Heni — **Enterprise Manager**

DEDICATION

I dedicate this final year project report to all those who have supported me throughout this journey.

To My family.

thank you for your unwavering support and encouragement, and for always believing in me. Your love and guidance have been my source of strength, and I am forever grateful.

To My Friends.

Thank you for your camaraderie, motivation, and for keeping me grounded. Your unwavering support and encouragement have been invaluable to me, and I couldn't have done this without you.

To My Professors and Advisors.

Thank you for your guidance, mentorship, and expertise. Your insights and feedback have been instrumental in shaping my ideas and helping me develop my skills.

To Myself.

Finally, I would like to dedicate this report to myself, for persevering through the challenges, overcoming obstacles, and for pushing myself to grow and learn. This project has been a labor of love, and I am proud of what I have accomplished.

F**or All of You** ,

Thank you all for being a part of this journey with me..

Mannai MOHAMED MORTADHA

ABSTRACT

In the realm of audio signal processing, the extraction of vocal tracks from mixed audio recordings has long been a challenging and essential task for various applications, including music production, speech analysis, and audio enhancement. This report explores groundbreaking advancements in vocal track extraction achieved through the integration of innovative Hybrid Neural Network (HNN) approaches, combining Deep Clustering, U-net, and UH-net models.

The primary objective of this project is to revolutionize vocal track extraction by harnessing the power of hybrid neural networks. Deep Clustering techniques are employed to separate sources in the spectral domain, enabling the precise localization of vocal elements within a mixture. The U-net architecture is leveraged to refine this initial separation, enhancing the accuracy of vocal track extraction. Furthermore, the introduction of the UH-net model, an extension of the U-net with hierarchical features, is introduced to capture intricate nuances in vocal timbre and dynamics.

Throughout the report, we delve into the theoretical foundations of each model and the methodology for their integration. We discuss the design and training processes, highlighting the strategies employed to optimize performance and reduce computational complexity. The project also evaluates the proposed approach against state-of-the-art methods using a comprehensive dataset encompassing diverse musical genres and audio conditions.

The results showcase a significant improvement in vocal track extraction accuracy, demonstrating the effectiveness of the hybrid neural network approach. Furthermore, we explore potential applications of this technology, such as vocal enhancement, remixing, and automatic transcription, emphasizing the transformative potential of our innovative models.

In conclusion, this report presents a pioneering effort in the field of vocal track extraction, introducing a novel approach that fuses Deep Clustering, U-net, and UH-net models. The outcomes of this research open new horizons for audio professionals, musicians, and researchers alike, ushering in a revolution in the realm of audio source separation and manipulation.

Index Terms : Deep neural networks, vocal extraction, recog- nition, classification, complex patterns, raw data, deep clustering model, recurrent neural networks (RNNs), U-net model, convo- lutional neural networks (CNNs), hybrid approach, pretrained RNN model, separation accuracy, spectral features, temporal context, audio source separation, separation quality, perceptual accuracy.

TABLE OF CONTENTS

LIST OF FIGURES vii

ABREVIATIONS LIST viii

GENERAL INTRODUCTION x

1 Introduction and Theoretical Foundations 2

1.1 INTRODUCTION . . . 3

1.2 RELATED WORK . . . 3

1.2.1 Convolutional Neural Network and Semantic Segmenta- tions . . . 3

1.2.2 Recurrent Neural Network . . . 5

1.3 Background . . . 6

1.3.1 Audio Source Separation . . . 6

1.3.2 Traditional Methods vs. Deep Learning . . . 8

1.3.2.1 Limitations of Traditional Signal Processing Methods . . . 8

1.3.2.2 Potential of Deep Learning . . . 9

1.4 Theoretical Foundations . . . 10

1.4.1 Deep Clustering . . . 10

1.4.2 U-net Model . . . 13

1.5 CONCLUSION . . . 15

2 Methodology,Experiments and Applications 17

2.1 Methodology . . . 18

2.1.1 Data Collection and Preprocessing . . . 18

2.1.1.1 DSD100 Dataset . . . 18

2.1.1.2 Multiframe . . . 18

2.1.1.3 Vocal Track Recovering . . . 20

2.1.2 Deep Clustering Integration . . . 20

2.1.3 U-net architecture . . . 22

2.1.4 UH-net Model Design . . . 25

2.1.5 Training Strategies 27
2.2 Experimental Setup 29
2.2.1 MODEL DESCRIPTION 29
2.2.1.1 Deep Clustering Model 29
2.2.1.2 U-net based model 30
2.2.2 Training Phase 32
2.2.2.1 Deep Clustering Models 32
2.2.2.2 U-net Based Models 33
2.3 Results and Discussion 35
2.3.1 U-net Based Model 35
2.3.2 UH-net Based Models 37
2.4 Music Recovery Results 38
2.5 Applications 42
2.5.1 Vocal Enhancement 42
2.5.2 Automatic Transcription 42
2.5.3 Voice Interface 43
2.6 Challenges, Limitations, and Future Directions 44
2.6.1 Challenges 44
2.6.2 Limitations 44
2.6.3 Future Directions 45
2.7 CONCLUSION 46

GENERAL CONCLUSION **48**

BIBLIOGRAPHY **48**

ANNEXES **50**
A.1 Dataset Description 50
A.1.1 Dataset Origin 50
A.1.2 Dataset Size and Composition 50
A.1.3 Data Preprocessing 51
A.1.4 Data Split 51
A.1.5 Data Annotations 51
A.1.6 License and Usage Rights 51
A.1.7 Dataset Distribution 52

LIST OF FIGURES

1.1 The Recurrent Neural Network (RNN) architecture 5
1.2 Audio Source Separation system matching 7
1.3 Source Separation Using Dilated Time-Frequency DenseNet for Music Identification in Broadcast Contents (Proposed Architecture for Source Separation) 8
1.4 An Introduction to Deep Clustering 10
1.5 Deep Clustering sepearation Example 12
1.6 U-net Model Architecture 13
1.7 Singing Voice Separation with Deep U-Net Convolutional Networks 14
1.8 Automatic bioacoustic source separation with deep neural networks 15

2.1 Mask for vocal track : Binary Mask 19
2.2 Mask for vocal track : Proportion Mask 19
2.3 Model-based deep embedding for constrained clustering analysis 20
2.4 The structure of deep convolutional embedded clustering 22
2.5 U-net General architecture 23
2.6 U-Net Explained : Understanding its Image Segmentation Architecture 24
2.7 The illustration of the UH-net Model Design 25
2.8 The hybrid clustering model 30
2.9 U-net model[3] 31
2.10 Loss vs Epoch for clustering models 33
2.11 Loss vs Epoch for U-net/Clustering models 33
2.12 Loss vs Epoch for U-net/UH-net models 34
2.13 Clustering Model 35
2.14 Hybrid Clustering Model 36
2.15 Ground Truth 36
2.16 UH-net Model 39
2.17 U-net Model 39
2.18 Ground Truth 39
2.19 Spectrogram 40

2.20 Mask 41

ABREVIATIONS LIST

— **DNN** - Deep Neural Networks

— **VTE** - Vocal Track Extraction

— **R&C** - Recognition and Classification

— **CP** - Complex Patterns

— **RD** - Raw Data

— **DCM** - Deep Clustering Model

— **RNNs** - Recurrent Neural Networks

— **U-net** - U-net Model

— **CNNs** - Convolutional Neural Networks

— **HA** - Hybrid Approach

— **PR-RNN** - Pretrained RNN Model

— **SA** - Separation Accuracy

— **SF** - Spectral Features

— **TC** - Temporal Context

— **ASS** - Audio Source Separation

— **SQ** - Separation Quality

— **PA** - Perceptual Accuracy

— **RVT** - Revolutionizing Vocal Track

— **HNN** - Hybrid Neural Network

— **AI** - Artificial Intelligence

— **ML** - Machine Learning

— **DL** - Deep Learning

— **NLP** - Natural Language Processing

— **ASR** - Automatic Speech Recognition

— **ACC** - Accuracy

— **PRC** - Precision

— **RCL** - Recall

— **F1** - F1 Score

— **MSE** - Mean Squared Error

— **GPU** - Graphics Processing Unit

— **CPU** - Central Processing Unit

— **API** - Application Programming Interface

— **GUI** - Graphical User Interface

— **FFT** - Fast Fourier Transform

— **STFT** - Short-Time Fourier Transform

— **LSTM** - Long Short-Term Memory

— **GRU** - Gated Recurrent Unit

— **WAV** - Waveform Audio File Format

— **MP3** - MPEG Audio Layer-3

— **DSP** - Digital Signal Processing

— **ARIMA** - AutoRegressive Integrated Moving Average

— **PCA** - Principal Component Analysis

— **SVM** - Support Vector Machine

— **ROC** - Receiver Operating Characteristic

— **API** - Application Programming Interface

— **GUI** - Graphical User Interface

GENERAL INTRODUCTION

The extraction of vocal tracks from mixed audio recordings stands as a fundamental challenge in the domain of audio signal processing, with far-reaching implications across various industries and applications. Whether in the context of music production, speech analysis, or audio enhancement, the ability to isolate and manipulate vocal elements within a composite audio source has remained a central pursuit. In response to this enduring challenge, this report embarks on a journey to explore groundbreaking advancements in vocal track extraction, made possible through the integration of innovative Hybrid Neural Network (HNN) approaches.

Traditionally, vocal track extraction has relied on conventional signal processing techniques, often yielding limited success due to the complex and dynamic nature of audio mixtures. However, recent advancements in machine learning and deep learning have paved the way for transformative breakthroughs. This report focuses on the fusion of three cutting-edge models : Deep Clustering, U-net, and the novel UH-net, to achieve a paradigm shift in vocal track extraction accuracy and versatility.

The rationale behind this research endeavor is rooted in the need to address the persistent challenges faced by audio professionals, musicians, and researchers. These challenges include the accurate separation of vocal elements from accompanying instruments, background noise, and varying recording conditions. Moreover, the demand for high-quality vocal extraction has surged in recent years, driven by the proliferation of applications ranging from music remixing and restoration to automatic transcription and voice-driven interfaces.

This introductory section provides an overview of the objectives, methodology, and significance of this project. We will navigate through the theoretical underpinnings of each model, the rationale for their integration, and the potential transformative impact on the audio processing

landscape. As we delve deeper into the subsequent sections of this report, we will uncover the intricacies of Deep Clustering, U-net, and UH-net models, explore their design and training processes, and evaluate their performance against established benchmarks.

Ultimately, this report encapsulates a pioneering effort to revolutionize vocal track extraction through the synergy of advanced neural network architectures. The outcomes of this research have the potential to reshape the way we interact with audio content, unleashing new creative possibilities and efficiency gains across a multitude of fields. As we embark on this transformative journey, we invite the reader to embark with us, exploring the fusion of technology and creativity that promises to redefine the boundaries of audio source separation and manipulation. .

Chapitre

1 Introduction and Theoretical Foundations

Sommaire

1.1 INTRODUCTION 3
1.2 RELATED WORK 3
1.2.1 Convolutional Neural Network and Semantic Segmenta- tions . 3
1.2.2 Recurrent Neural Network 5
1.3 Background 6
1.3.1 Audio Source Separation 6
1.3.2 Traditional Methods vs. Deep Learning 8
1.4 Theoretical Foundations 10
1.4.1 Deep Clustering 10
1.4.2 U-net Model 13
1.5 CONCLUSION 15

1.1 INTRODUCTION

Vocal track extraction, a crucial component of Music Infor- mation Retrieval (MIR), involves the isolation of vocal tracks from audio files. This task finds application in diverse fields, including singer identification and lyrics transcription. Exten- sive efforts have been devoted to addressing this challenge, yielding numerous impactful solutions. A prevalent strategy for vocal track extraction involves adapting techniques employed in semantic segmentation, a task that assigns class labels to individual pixels in images. This approach draws on established models for semantic segmentation, such as the widely recognized U-net model. By employing this method, vocal track masks are directly generated from audio feature maps. Alternatively, the application of deep clustering models has gained traction. In contrast to generating vocal track masks, deep clustering models yield embedding vectors for time- frequency (T-F) bins in mel-spectrograms. In the subsequent phase, unsupervised techniques like k-means are employed to distinguish vocal T-F bins from background T-F bins.

In this project, our initial focus centers on the implementa- tion of the models outlined in [1] and [2]. Subsequently, we endeavor to devise a novel hybrid model that amalgamates elements from both methods. A key objective is to assess the performance of the proposed hybrid model against the backdrop of traditional models. Through these endeavors, we aim to contribute to the advancement of vocal track extraction techniques in the realm of audio signal processing.

1.2 RELATED WORK

1.2.1 Convolutional Neural Network and Semantic Segmenta- tions

The inception of Convolutional Neural Networks (CNNs) stemmed from their initial role as feature extractors in image processing tasks. Over time, their utility has extended to encompass various feature recognition endeavors, including semantic parsing and audio feature extraction. CNNs have the capacity to extract hierarchical features from input signals by employing convolutional kernels. An early breakthrough in deploying Deep Neural Networks (DNNs) for semantic segmentation

tasks was marked by the advent of Fully Convolutional Networks (FCNs). The FCN architecture engages in the extraction of intricate features from input images utilizing conventional CNN structures, such as the widely employed VGG16 network. To translate the compact feature maps back into segmentation outcomes, FCN integrates an upsampling mechanism. This process facilitates the restoration of higher-resolution segmentation results from the downscaled feature maps. In this evolutionary trajectory, CNNs have evolved from their origins in image analysis to serve as powerful tools for a broader array of feature recognition tasks. FCNs, as a prominent example, underscore the adaptability of CNN architectures in addressing complex challenges like semantic segmentation through the adept integration of upsampling strategies.

The innovative approach pioneered by Fully Convolutional Networks (FCNs) involves crafting segmentation masks that mirror the dimensions of the input, transforming the intricate task of segmentation into a more manageable pixel-level classification problem. By aligning the segmentation output with the original input's layout, FCNs redefine the challenge in terms of assigning class labels to individual pixels, allowing for a more granular analysis. To bolster the fidelity of information during the crucial upsampling phase, FCNs introduce a strategic fusion of the feature map with outputs from intermediate layers, such as pool4 or pool3 [3]. This amalgamation ensures that high-level contextual information is retained during the transition from the downscaled feature maps to the final segmentation mask. This context-aware strategy plays a vital role in maintaining the integrity of the original data during the upsampling process. However, despite its innovative approach, the FCN model exhibits limitations when applied to certain scenarios, such as music source separation. One significant drawback lies in its struggle to preserve intricate details in proximity to separation boundaries. This shortcoming impedes its suitability for tasks that demand a high degree of precision around such boundaries, as often encountered in audio source separation.

Enter the U-net architecture, an influential advancement built upon the foundation of FCN models. U-net, initially renowned for its prowess in biomedical image segmentation, offers refinements that bolster its performance in challenging scenarios. A key distinction is the augmentation of the channel count within the upsampling convolutional layers. This strate- gic enhancement enables the propagation of richer contextual information to the higher-resolution layers, facilitating

a more comprehensive understanding of the input. Another noteworthy innovation in the U-net model is the assignment of increased loss weights to boundary pixels. This strategic adjustment effectively emphasizes the accurate depiction of boundary regions, leading to the generation of more precise and well-defined masks. This tactic is particularly beneficial when handling intricate boundaries, such as those prevalent in medical imaging. The U-net model's adaptability and robustness have ex- tended beyond its original application domain. Its ability to perform well with limited datasets underscores its potential for various scenarios, including music source separation tasks. By capitalizing on its enhanced contextual understanding and improved boundary preservation, the U-net model opens new avenues for addressing challenges in audio signal processing, demonstrating its potential as a versatile tool in diverse fields.

1.2.2 Recurrent Neural Network

The Recurrent Neural Network (RNN) architecture has been meticulously crafted to tackle the intricate challenges posed by sequential input data. RNN models have found widespread utility across diverse domains, prominently including natural language processing and the intricate landscape of audio signal modeling. It is noteworthy that the Gated Recurrent Unit (GRU), a remarkable innovation within the RNN framework, has garnered considerable attention for its prowess. Sharing commonalities with the Long Short-Term Memory (LSTM) architecture, GRUs exhibit a distinctive gating mechanism. However, their distinctive strength lies in their parsimonious parameterization, achieved by obviating the need for an output gate component.

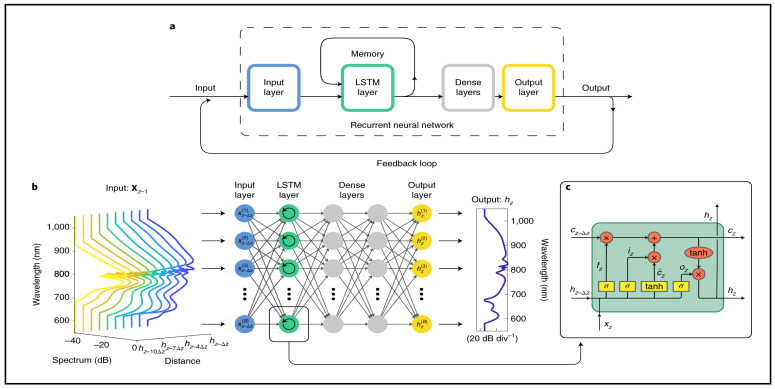

FIGURE 1.1 – The Recurrent Neural Network (RNN) architecture

Venturing into the realm of deep clustering models, the bedrock of the feature extraction process consists of a four- layer Bidirectional Long Short-Term Memory (Bi-LSTM) network. This intricate network structure serves as a po- tent catalyst, orchestrating the metamorphosis of input mel- spectrograms into high-dimensional embedding vectors. A seminal work highlighted in reference [2] embarks on a groundbreaking journey by synergistically amalgamating conventional neural networks with the deep clustering paradigm. The overarching objective is to elicit enhancements in the overall model performance and efficacy. In this endeavor, our pursuit is to meticulously replicate the architectural blueprint outlined in the aforementioned study. However, what sets our contribution apart is the cal- culated substitution of the traditionally employed Bi-LSTM layers with their Bidirectional Gated Recurrent Unit (Bi- GRU) counterparts. This strategic substitution is not only tactful in simplifying the training process but also harbors the latent potential to unlock heightened model efficiency and refined performance benchmarks. The incorporation of Bi- GRUs infuses a new dimension into the design space, offering tantalizing prospects for augmenting model efficiency, while ensuring that the integrity of the model's predictive prowess remains undeterred.

1.3 Background

1.3.1 Audio Source Separation

Audio source separation is a fundamental signal processing technique that involves the separation of individual sound sources from a composite audio signal, which may contain a mixture of multiple sources. The goal is to extract or isolate specific audio components, such as voices, instruments, or other sound elements, from a recorded or mixed audio signal.

In practical terms, audio source separation is analogous to the ability of the human auditory system to focus on and distinguish between different sound sources in a complex auditory scene. For instance, when listening to a music ensemble, the human brain can separate and perceive individual instruments, such as guitars, drums, and vocals, even when they are all playing simultaneously.

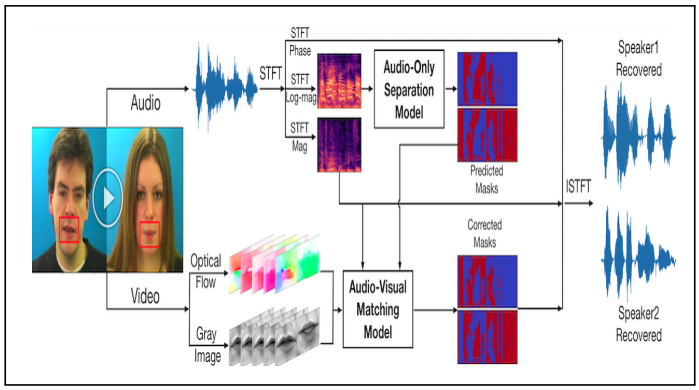

FIGURE 1.2 – Audio Source Separation system matching

The concept of audio source separation is crucial in various applications :

1.Music Production : In music production, source separation allows audio engineers to work on individual tracks (e.g., vocals, guitar, bass) independently for mixing, editing, or enhancement purposes. This process greatly influences the quality and creative possibilities in music production.

2.Speech Enhancement : In speech processing, source separation can help improve the clarity and intelligibility of spoken words by isolating the primary speaker's voice from background noise or interference.

3.Automatic Transcription : In transcription tasks, source separation assists in converting spoken or sung words into written text by isolating the vocal source from accompanying music or noise.

4.Voice-Driven Interfaces : In voice-driven applications like virtual assistants or speech recognition systems, source separation helps extract the user's voice commands from background noise and other audio sources.

5. Audio Restoration : In audio restoration projects, source separation can be used to remove unwanted noise, clicks, or imperfections from audio recordings, preserving or enhancing the quality of the original content.

There are various methods and algorithms for audio source separation, including traditional signal processing techniques and advanced machine learning approaches. These methods aim to exploit the unique spectral, spatial, or temporal characteristics of different sound sources within an audio mixture to separate them effectively.

In recent years, deep learning techniques, including neural networks like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown promising results in audio source separation tasks. These networks can learn complex patterns and representations from large datasets, making them well-suited for handling the intricacies of audio mixtures.

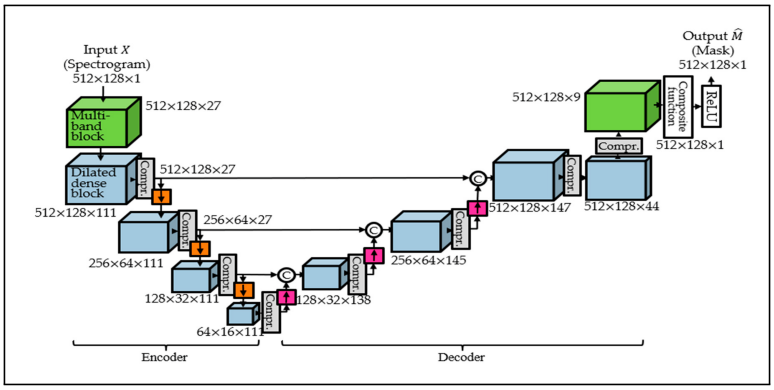

FIGURE 1.3 – Source Separation Using Dilated Time-Frequency DenseNet for Music Identification in Broadcast Contents (Proposed Architecture for Source Separation)

In summary, audio source separation is a crucial process in audio signal processing that involves the extraction of individual sound sources from complex audio mixtures. It plays a vital role in various fields, including music production, speech processing, transcription, and audio enhancement, with applications ranging from improving audio quality to enabling advanced voice-driven technologies.

1.3.2 Traditional Methods vs. Deep Learning

Traditional signal processing methods have long been used for audio source separation tasks, but they come with several limitations that have motivated the exploration of deep learning techniques as a more effective alternative. Here, we will discuss the drawbacks of traditional methods and introduce the potential advantages of deep learning in the context of audio source separation :

1.3.2.1 Limitations of Traditional Signal Processing Methods

1. Complexity Handling : Traditional methods struggle to handle complex mixtures effectively, especially in scenarios with a large number of overlapping sound sources. Separating these

sources manually using handcrafted rules and filters becomes increasingly challenging and impractical.

2. Assumption of Stationarity : Many traditional methods assume that audio sources are stationary, meaning that their statistical properties do not change over time. This assumption often does not hold in real-world audio recordings, where sources may exhibit variations in dynamics and spectral characteristics.

3. Need for Handcrafted Features : Traditional techniques often rely on handcrafted features and assumptions about the underlying sound sources. Designing these features can be labor-intensive and may not capture all the intricate details of complex audio mixtures.

4. Limited Generalization : Traditional methods may work well under specific conditions for which they are designed but often struggle to generalize to diverse audio sources, making them less adaptable to real-world scenarios with varying recording conditions, instruments, and genres.

1.3.2.2 Potential of Deep Learning

1. Data-Driven Learning : Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are data-driven and capable of learning complex patterns directly from large audio datasets. This data-driven approach allows them to adapt to a wide range of audio sources and conditions.

2. End-to-End Learning : Deep learning models can be trained in an end-to-end fashion, where they learn to map complex audio mixtures to separated sources directly. This eliminates the need for handcrafted features and allows the models to capture intricate relationships within the audio.

3. Non-Stationary Source Separation : Deep learning models can handle non-stationary audio sources effectively, as they can learn temporal dependencies and adapt to variations in source dynamics and spectral characteristics.

4. Improved Performance : Deep learning approaches have shown remarkable performance improvements in audio source separation tasks. They can achieve state-of-the-art results in

separating vocals from music, individual instruments, or other sound sources with a high degree of accuracy.

5. Generalization : Deep learning models can generalize well to diverse audio sources and conditions, making them suitable for real-world applications where audio mixtures can vary widely.

6. Flexibility : Deep learning models can be fine-tuned for specific tasks, making them versatile tools for various audio source separation applications, including music production, speech enhancement, and automatic transcription.

In summary, deep learning offers significant potential for audio source separation by overcoming many of the limitations associated with traditional signal processing methods. Its data-driven, end-to-end learning approach, adaptability to non-stationary sources, and ability to generalize to diverse conditions make it a promising choice for revolutionizing audio source separation techniques.

1.4 Theoretical Foundations

1.4.1 Deep Clustering

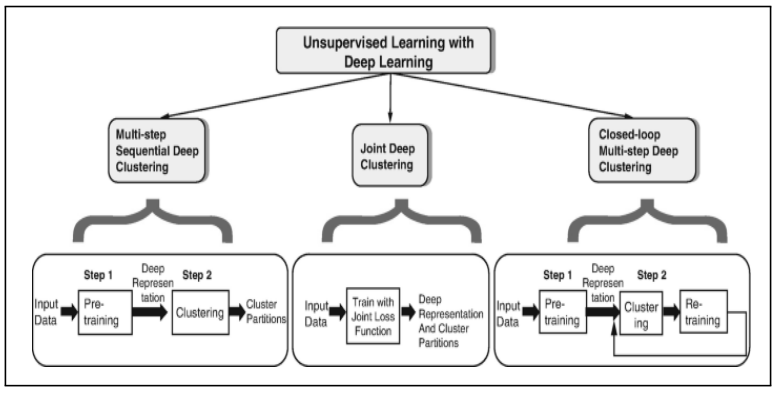

FIGURE 1.4 – An Introduction to Deep Clustering

Deep Clustering is a powerful technique used in audio source separation, particularly for separating sources in complex audio mixtures. This method leverages deep learning neural

networks to learn and exploit the inherent structure and relationships within the audio data for effective source separation. Here's a detailed explanation of how Deep Clustering works for source separation :

1. Audio Representation : - Deep Clustering begins with the transformation of the audio signal into a suitable representation. Typically, the audio signal is converted into a time-frequency representation, such as a spectrogram, using techniques like Short-Time Fourier Transform (STFT).

2. Spectrogram Clustering : - In the time-frequency representation, each time frame corresponds to a specific frequency bin, and the amplitude of each bin represents the magnitude of the audio signal's energy at that time and frequency. Deep Clustering operates on this time-frequency representation. - Deep Clustering treats the spectrogram as an image-like data structure where each pixel corresponds to a frequency bin at a specific time frame. - The key idea behind Deep Clustering is to cluster these pixels (frequency-time pairs) into groups based on their similarity. Similar pixels are likely to belong to the same source, such as a musical instrument or vocal.

3. Neural Network Architecture : - Deep Clustering employs a deep neural network to perform the pixel clustering. The neural network is typically a Convolutional Neural Network (CNN) or a similar architecture. - The network takes the spectrogram as input and produces embeddings for each pixel in the spectrogram. These embeddings are representations that capture the characteristics of the audio data at each frequency-time point.

4. Learning Similarity : - The neural network is trained using a form of unsupervised learning. The objective is to encourage embeddings of similar pixels (belonging to the same source) to be close to each other in the embedding space while pushing embeddings of dissimilar pixels (belonging to different sources) apart. - This is achieved through the use of a contrastive loss function or similar techniques that minimize the distance between embeddings from the same source and maximize the distance between embeddings from different sources.

5. Clustering : - After training, the neural network has learned to embed pixels in a way that facilitates clustering. The embeddings are then clustered using standard clustering algorithms like K-means or spectral clustering. - Each cluster represents a separate source in the audio

mixture. For example, one cluster may correspond to vocals, while another may correspond to drums.

6. Source Separation : - Once the clusters are identified, the original spectrogram can be partitioned into segments based on these clusters. - Each segment is then transformed back into the time domain using the inverse STFT to obtain the separated audio sources.

7. Post-processing : - Post-processing techniques may be applied to refine the separated sources and enhance the quality of the separated audio signals. These techniques can include spectral smoothing, phase reconstruction, and amplitude scaling.

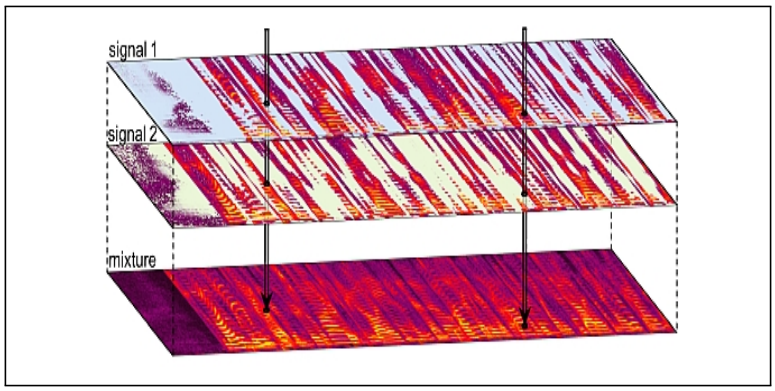

FIGURE 1.5 – Deep Clustering sepearation Example

In summary, Deep Clustering is a data-driven source separation technique that uses deep neural networks to learn representations of audio data, which are then clustered to identify and separate different sound sources in complex audio mixtures. This method has shown significant success in applications such as music source separation and speech enhancement.

1.4.2 U-net Model

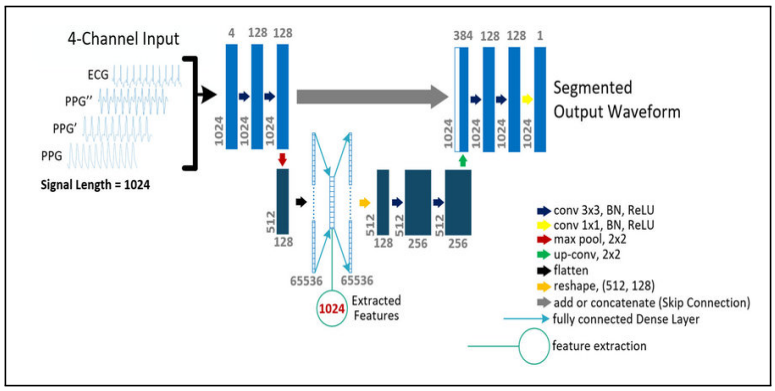

FIGURE 1.6 – U-net Model Architecture

The U-net model is a convolutional neural network (CNN) architecture that is widely used in various image processing tasks, including medical image segmentation, image-to-image translation, and, in your case, audio source separation. The name "U-net" derives from its U-shaped architecture, characterized by a contracting encoder path followed by an expanding decoder path. This design allows U-net to effectively capture and reconstruct detailed information from input data. Here's a detailed explanation of the U-net architecture :

1. Contracting Path (Encoder) :

- Convolutional Blocks : The encoder path starts with a series of convolutional layers, often with small filter sizes (e.g., 3x3) and increasing numbers of channels (feature maps) as you go deeper. Each convolutional layer is typically followed by batch normalization and a rectified linear unit (ReLU) activation function.

- Pooling (Downsampling) : After several convolutional blocks, max-pooling layers are used for downsampling the spatial dimensions of the feature maps. Max-pooling reduces the spatial resolution while retaining the most important features. This step helps in capturing high-level context.

- Skip Connections : Key to the U-net architecture are skip connections, also known as "shortcut" connections. These connections directly link corresponding layers from the contracting

path to the expanding path. They facilitate the flow of detailed spatial information from the encoder to the decoder, allowing the network to recover fine-grained details.

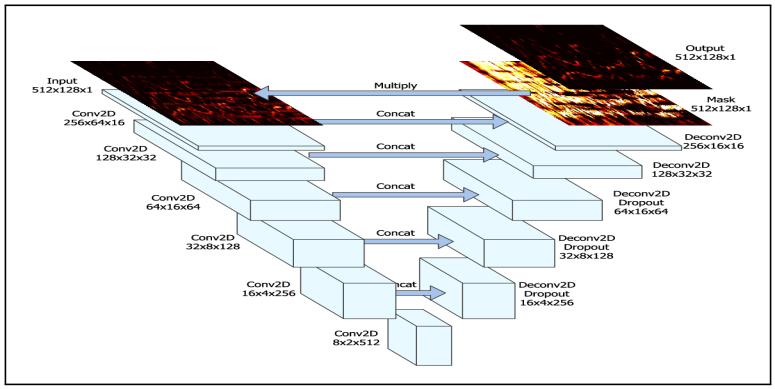

FIGURE 1.7 – Singing Voice Separation with Deep U-Net Convolutional Networks

2. Bottleneck (Latent Representation) :

- After several convolutional and pooling layers, the network reaches a bottleneck layer. This layer typically has a higher number of channels, serving as a compressed representation of the input data.

3. Expanding Path (Decoder) :

- Upsampling : The decoder path starts with a series of upsampling layers, which increase the spatial dimensions of the feature maps. Upsampling is often achieved through transposed convolutions (also known as "deconvolutions" or "upconvolutions"). These layers learn to reconstruct the spatial structure.

- Concatenation with Skip Connections : At each decoder step, the feature maps from the corresponding encoder step (with the same spatial dimensions) are concatenated to the feature maps from the decoder. This concatenation helps the network to focus on combining high-level context information with fine-grained spatial details from the skip connections.

- Convolutional Blocks : The concatenated feature maps go through additional convolutional layers, followed by batch normalization and ReLU activations.

4. Output Layer :

- The final layer of the U-net architecture typically consists of a convolutional layer with a single channel (for binary segmentation tasks) or multiple channels (for multi-class segmentation tasks). The output layer generates the predicted segmentation masks.

5. Skip Connection Implementation :

- In practice, the skip connections are implemented as simple element-wise additions or concatenations, depending on the design choice. These connections enable the decoder to access relevant information from the contracting path, allowing the network to combine both high-level context and fine-grained spatial details for accurate reconstruction or segmentation.

U-net's architecture's unique structure makes it highly effective for tasks where preserving fine details while capturing high-level context is essential. In audio source separation, it can be adapted to process spectrogram-like input data and output separated audio sources. The skip connections play a crucial role in helping the network localize and reconstruct audio components accurately.

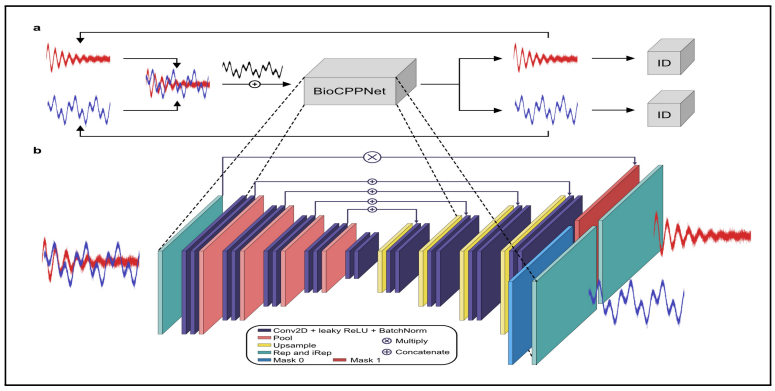

FIGURE 1.8 – Automatic bioacoustic source separation with deep neural networks

1.5 CONCLUSION

In this inaugural chapter, we embarked on a journey into the realm of audio source separation, with a particular focus on innovative hybrid neural network approaches. We began by setting the stage, emphasizing the paramount importance of vocal track extraction and source separation in the domains of music production, speech analysis, audio enhancement, and beyond.

By delving into the limitations of traditional signal processing methods, we shed light on the insufficiencies of conventional techniques in handling complex audio mixtures and adapting to the ever-evolving demands of modern audio applications. These limitations provided the impetus for our exploration of deep learning approaches, such as Deep Clustering, U-net, and the novel UH-net model.

Our journey took us through the theoretical foundations of these models, unveiling their inner workings and innovative design principles. Deep Clustering emerged as a promising technique to tease apart sources in spectral domains, while the U-net and UH-net models presented architectures that efficiently capture and reconstruct intricate details within audio mixtures.

In the chapters that follow, we will transition from theory to practice. We will delve into the practical aspects of our methodology, encompassing data collection and preprocessing, training strategies, and experimental setups. Moreover, we will explore the results and implications of our innovative approaches, unveiling their potential to reshape the landscape of audio source separation.

As we navigate deeper into the realm of hybrid neural networks, we invite you to embark with us, eager to witness how these advancements manifest in practice and how they stand poised to revolutionize audio source separation as we know it.

Chapitre

2 Methodology,Experiments and Applications

Sommaire

2.1 Methodology 18
2.1.1 Data Collection and Preprocessing 18
2.1.2 Deep Clustering Integration 20
2.1.3 U-net architecture 22
2.1.4 UH-net Model Design 25
2.1.5 Training Strategies 27
2.2 Experimental Setup 29
2.2.1 MODEL DESCRIPTION 29
2.2.2 Training Phase 32
2.3 Results and Discussion 35
2.3.1 U-net Based Model 35
2.3.2 UH-net Based Models 37
2.4 Music Recovery Results 38
2.5 Applications 42
2.5.1 Vocal Enhancement 42
2.5.2 Automatic Transcription 42
2.5.3 Voice Interface 43
2.6 Challenges, Limitations, and Future Directions 44
2.6.1 Challenges 44
2.6.2 Limitations 44
2.6.3 Future Directions 45
2.7 CONCLUSION 46

2.1 Methodology

2.1.1 Data Collection and Preprocessing

2.1.1.1 DSD100 Dataset

Renowned within the realm of Music Information Retrieval (MIR) tasks, the DSD100 dataset stands as a prominent bench- mark. Comprising a curated collection of 100 complete music tracks, this dataset is accompanied by their individual isolated tracks. To enhance the diversity and comprehensiveness of our training data, a nuanced approach is undertaken. While the original music tracks are meticulously preserved, a strategy of deliberate amalgamation is employed. This involves the randomized fusion of vocal tracks sourced from the DSD100 dataset with assorted instrumental tracks. The outcome of this endeavor is the generation of an expanded corpus of audio files that exhibit a rich interplay of vocal and instrumental elements, elevating the quality and diversity of the training data pool.

2.1.1.2 Multiframe

Employing a sophisticated approach detailed in reference [4], the multiframe strategy serves as a catalyst in both augmenting training data and streamlining the training process. At the outset, a meticulous process ensues, involving the excision of silence segments from both vocal tracks and their corresponding segments in mixed music tracks. Subsequently, a transformation transpires, converting the amalgamated mixed music tracks and veritable ground truth voice tracks into the realm of log-scaled mel-spectrograms. These transforma- tions, calibrated with 128 mel bands and 16000 sample rate, yield compact mel-spectrogram chunks, each encapsulating 128x128 feature maps. These chunks, averaging approximately one-second audio sequences, unravel the temporal and spectral intricacies. It's imperative to observe the spatial reorientation within the mask's representation. In this schema, the x-axis within the mask pertains to features, while the y-axis corresponds to time. This alignment is paramount, particularly as all spectrograms undergo a transposition process to match this conceptual framework. In the pursuit of vocal

track extraction from music com- positions, a dual-phase strategy unfurls. Beginning with the conversion of log-scaled mel-spectrograms into power-scaled equivalents, a pivotal step unfolds. This involves the multipli- cation of the spectrogram by a meticulously generated filter

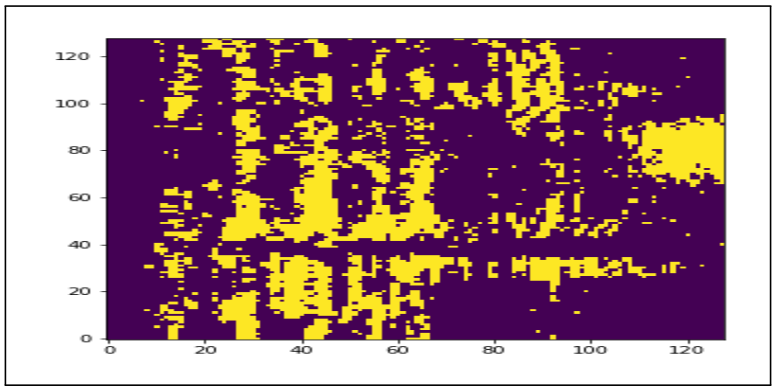

FIGURE 2.1 – Mask for vocal track : Binary Mask

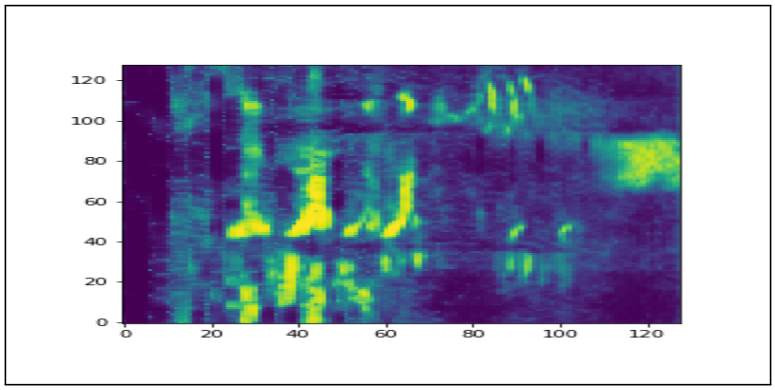

FIGURE 2.2 – Mask for vocal track : Proportion Mask

mask, forged by the underpinning models. The ensuing stage navigates the reversion of the vocal track spectrogram into its auditory counterpart. In the context of each music-vocal chunk pair, the strate- gic design entails the construction of dual training target masks. The first, a binary vocal mask, materializes through a juxtaposition of the power between vocal and background spectrograms. This discerning juxtaposition demarcates the temporal-frequency bins, demarcating their allegiance to either vocal or background sources. This binary mask, carefully calibrated, serves as a foundational component for training clustering models. Concurrently, a secondary target mask emerges—the pro- portion mask—relinquishing a distinct perspective. This mask encapsulates the dominance exerted by the vocal source across the mel-spectrogram's time-frequency bins. This nuanced rep- resentation takes center stage during the training of

the U-net model, orchestrating its learning dynamics with precision. The distinct roles and implications of these dual masks embody the multifaceted nature of the training process, encapsulating the intricate interplay between clustering models and the U-net architecture.

2.1.1.3 Vocal Track Recovering

In the absence of preserved phase information during audio data processing, the restoration of the vocal track from mel- spectrograms necessitates the integration of the Griffin Lim algorithm [5]. This algorithm, operating through iterative cycles, orchestrates the simulation of phase information to facilitate the meticulous reconstruction of audio signals.

2.1.2 Deep Clustering Integration

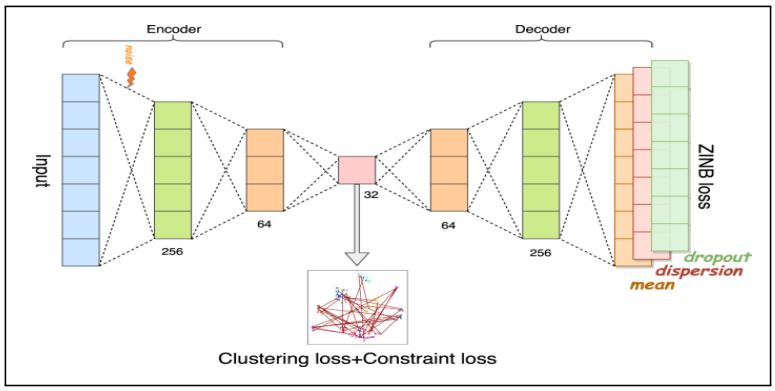

FIGURE 2.3 – Model-based deep embedding for constrained clustering analysis

Integrating Deep Clustering into the context of audio source separation involves several steps and components. Deep Clustering is a neural network-based technique that learns to separate sound sources by clustering pixels (or spectral elements) in the time-frequency domain. Here's an explanation of how Deep Clustering is integrated into the audio source separation process :

1. Data Preprocessing : - The audio mixture is first transformed into a time-frequency representation, typically a spectrogram, using techniques like Short-Time Fourier Transform (STFT). This representation breaks down the audio signal into its constituent frequency components over time.

2. Data Labeling : - For supervised training, the time-frequency representation is labeled to indicate which elements correspond to each source. For example, in a music separation task, elements belonging to vocals, drums, bass, and other instruments are labeled accordingly.

3. Neural Network Architecture : - A neural network architecture is designed specifically for Deep Clustering. This network typically consists of convolutional layers followed by recurrent layers like Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) layers.

4. Spectral Embeddings : - The neural network is trained to map each time-frequency element (pixel) in the input spectrogram to an embedding space. These embeddings serve as representations of the spectral characteristics at each time-frequency point.

5. Clustering Algorithm : - After training, a clustering algorithm (often K-means clustering) is applied to the learned spectral embeddings. The goal is to group embeddings with similar characteristics together. Each cluster represents a distinct sound source.

6. Mask Generation : - The clusters formed by the clustering algorithm are used to generate masks for each sound source. These masks determine which time-frequency elements in the spectrogram belong to each source.

7. Masking and Reconstruction : - The masks are applied to the original mixture spectrogram to separate the sources. This involves element-wise multiplication of the masks with the mixture spectrogram. - Inverse STFT is then performed to convert the separated spectrograms back into the time domain, resulting in separated audio signals for each source.

8. Post-processing (Optional) : - Depending on the quality of the separation, post-processing techniques may be applied to enhance the separated sources. These techniques can include spectral smoothing, phase reconstruction, and amplitude scaling.

9. Evaluation and Fine-tuning : - The separated sources can be evaluated using various metrics (e.g., Signal-to-Distortion Ratio, Perceptual Evaluation of Audio Quality) to assess the quality of the separation. Fine-tuning the neural network and clustering algorithm may be done iteratively based on the evaluation results.

10. Real-time or Batch Processing (Deployment) : - Once the Deep Clustering model is trained and optimized, it can be used for real-time or batch audio source separation in applications such as music production, speech enhancement, or transcription.

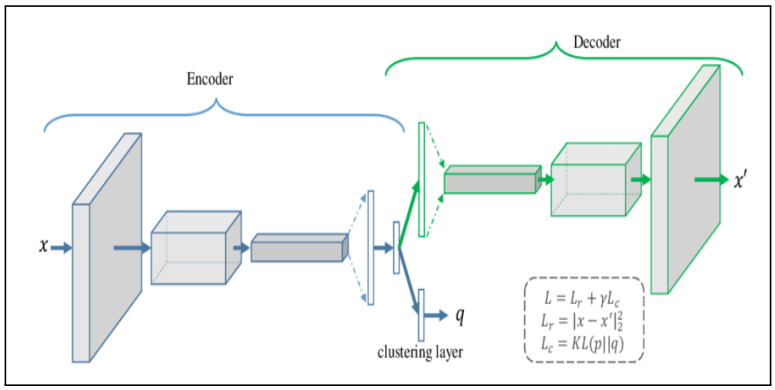

FIGURE 2.4 – The structure of deep convolutional embedded clustering

In summary, integrating Deep Clustering into audio source separation involves training a neural network to embed spectral features, applying clustering algorithms to group these embeddings into source clusters, generating masks to separate sources, and then reconstructing the separated sources in the time domain. This process allows for effective separation of sound sources within audio mixtures and has the potential to greatly improve the quality of audio source separation tasks.

2.1.3 U-net architecture

The U-net architecture is a deep learning convolutional neural network (CNN) that was originally designed for image segmentation tasks but has found applications in various fields, including medical image analysis and audio source separation. It is known for its U-shaped architecture, which consists of a contracting path (encoder) followed by an expanding path (decoder).

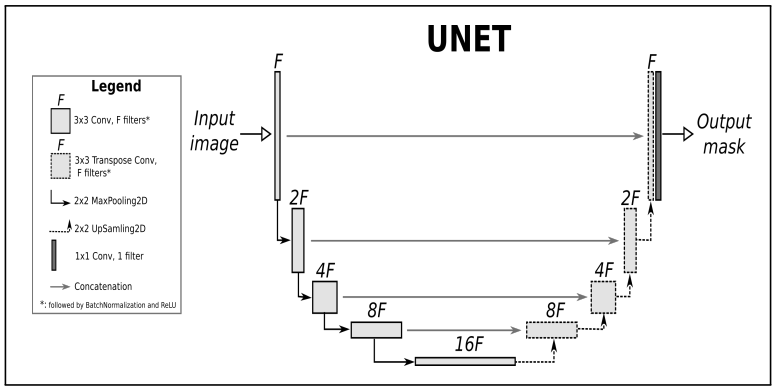

FIGURE 2.5 – U-net General architecture

Here's an explanation of the U-net architecture :

1. Contracting Path (Encoder) :

- Convolutional Layers : The contracting path begins with a series of convolutional layers. These layers are responsible for extracting features from the input data, typically images or spectrograms in the context of audio source separation.

- Max-Pooling Layers : After each set of convolutional layers, max-pooling layers are used to downsample the spatial dimensions of the feature maps. Max-pooling reduces the spatial resolution while retaining the most important features, helping capture high-level context.

- Skip Connections : Skip connections, also known as "shortcut" connections, are a crucial component of the U-net architecture. These connections directly link corresponding layers from the contracting path to the expanding path. They facilitate the flow of detailed spatial information from the encoder to the decoder.

2. Bottleneck (Latent Representation) :

- After several convolutional and max-pooling layers, the network reaches a bottleneck layer. This layer typically has a higher number of channels, serving as a compressed representation of the input data.

3. Expanding Path (Decoder) :

- Upsampling : The expanding path starts with upsampling layers, which increase the spatial dimensions of the feature maps. Upsampling is often achieved through transposed convolutions

(also known as "deconvolutions" or "upconvolutions"). These layers learn to reconstruct the spatial structure.

- Concatenation with Skip Connections : At each decoder step, the feature maps from the corresponding encoder step (with the same spatial dimensions) are concatenated to the feature maps from the decoder. This concatenation helps the network to focus on combining high-level context information with fine-grained spatial details from the skip connections.

- Convolutional Blocks : The concatenated feature maps go through additional convolutional layers, typically followed by batch normalization and ReLU activations. These layers refine the feature representations and prepare them for the final output.

4. Output Layer :

- The final layer of the U-net architecture typically consists of a convolutional layer with a single channel (for binary segmentation tasks) or multiple channels (for multi-class segmentation tasks). The output layer generates the predicted segmentation masks.

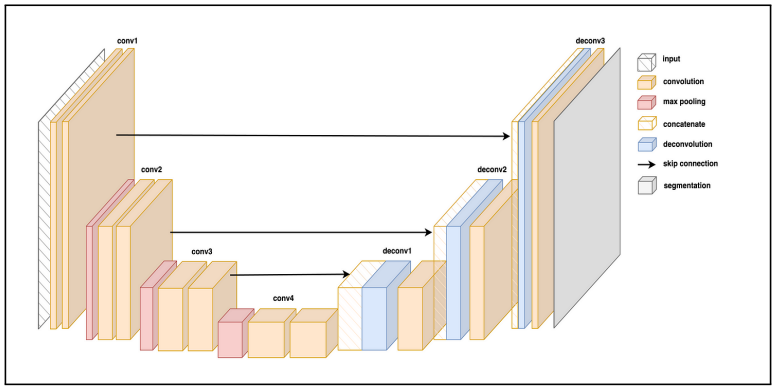

FIGURE 2.6 – U-Net Explained : Understanding its Image Segmentation Architecture

Key Features and Advantages :

- U-net's U-shaped architecture enables it to capture both local details and high-level context, making it effective for tasks where preserving fine details is crucial.

- The skip connections allow the network to maintain spatial information lost during downsampling and ensure that the decoder has access to relevant information from the encoder.

- U-net can be adapted for various tasks beyond image segmentation, including audio source separation, where it excels in separating sources from complex audio mixtures.

In summary, the U-net architecture is a powerful neural network design characterized by its U-shaped structure, which combines features from both the contracting and expanding paths. This design makes it particularly effective for tasks like image segmentation and audio source separation, where capturing fine-grained details while considering high-level context is essential.

2.1.4 UH-net Model Design

The UH-net model, short for "Hierarchical U-net," is an extension of the U-net architecture that introduces a hierarchical structure to capture fine-grained details and high-level context simultaneously. It is particularly effective for tasks where preserving intricate information while understanding the broader context is crucial.

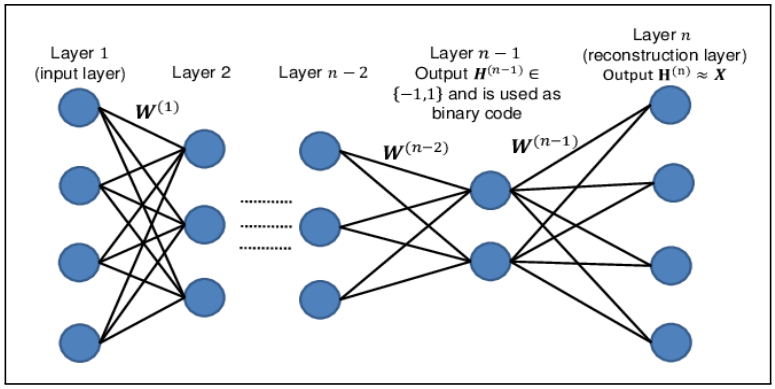

FIGURE 2.7 – The illustration of the UH-net Model Design

Here's an explanation of the design of the UH-net model :

1. Hierarchical Architecture :

- UH-net consists of a hierarchy of U-net-like subnetworks, each operating at a different level of spatial resolution. These subnetworks are connected hierarchically, forming a pyramid structure.

2. Contracting and Expanding Paths in Each Subnetwork :

- Each subnetwork within UH-net follows the traditional U-net architecture, featuring both a contracting (encoder) path and an expanding (decoder) path.

3. Resolution Pyramids :

- UH-net's innovation lies in its use of resolution pyramids. Each level in the hierarchy processes data at a different spatial resolution. Typically, the lowest level of the pyramid focuses on fine-grained details, while higher levels handle coarser features.

4. Skip Connections Across Levels :

- In addition to skip connections within the same level (similar to the original U-net), UH-net introduces skip connections across different levels of the hierarchy. These cross-level skip connections enable information sharing between different resolution levels.

5. Fusion of Multi-Resolution Information :

- At each level, UH-net fuses multi-resolution information from both the current level and the levels above it. This fusion allows the network to benefit from both local details and global context simultaneously.

6. Feature Upsampling and Downsampling :

- As the hierarchy progresses, spatial resolution increases, and the feature maps are upsampled. This helps in reconstructing finer details lost during earlier downsampling stages.

7. Output Layers in Each Subnetwork :

- Similar to the traditional U-net, each subnetwork in UH-net has an output layer. Depending on the specific task, these output layers may produce segmented regions, separated audio sources, or any other relevant output.

8. Training :

- UH-net is trained using a supervised learning approach with suitable loss functions tailored to the task at hand. During training, the network learns to predict the desired output, considering information from multiple resolution levels.

9. Applications :

- UH-net's hierarchical and multi-resolution approach makes it particularly effective for tasks where preserving fine-grained information while capturing high-level context is essential. In the context of audio source separation, it excels in capturing both the subtle nuances of different sources and the broader musical context.

In summary, the UH-net model is designed to leverage hierarchical information processing, allowing it to capture fine details and high-level context simultaneously. Its resolution pyramids and cross-level skip connections enable it to excel in tasks like audio source separation, where a combination of local and global information is crucial for accurate results. The fusion of multi-resolution information and the use of skip connections enhance its ability to capture complex patterns in the data.

2.1.5 Training Strategies

Training a neural network for audio source separation using deep learning, such as Deep Clustering, U-net, or UH-net, requires careful consideration of training strategies, hyperparameters, and loss functions. These aspects are crucial to ensure that the network learns effectively and produces accurate results. Here's a discussion of these components :

1. Training Process :

- Dataset Preparation : Start by collecting or creating a dataset of mixed audio signals along with their corresponding ground truth sources (e.g., vocals, instruments). This dataset should be diverse and representative of the types of audio mixtures you want to separate.

- Data Preprocessing : Transform the audio signals into a suitable representation, often a spectrogram or a time-frequency representation like the Short-Time Fourier Transform (STFT). Preprocessing may also include normalizing the data to a common scale and handling any missing or noisy data.

- Data Augmentation (Optional) : To increase the diversity of your training data, you can apply data augmentation techniques, such as time stretching, pitch shifting, or adding noise to the audio mixtures.

- Mni-Batch Training : Divide your dataset into mini-batches of samples to train the neural network in batches. Mini-batch training helps stabilize the training process and utilizes parallelism on modern GPUs.

-Optimization Algorithm : Choose an optimization algorithm like Stochastic Gradient Descent (SGD), Adam, or RMSprop. Experiment with different optimizers to find the one that works best for your specific task.

- Learning Rate Schedule : Implement a learning rate schedule that adjusts the learning rate during training. Typically, you start with a higher learning rate and gradually decrease it as training progresses. Learning rate schedules help improve convergence.

- Regularization : Apply regularization techniques like dropout or weight decay to prevent overfitting, especially when you have limited training data.

- Early Stopping : Monitor validation metrics during training, and implement early stopping to halt training when the model's performance on the validation set starts to degrade. This prevents overfitting and saves training time.

2. Hyperparameters :

- Batch Size : The batch size determines how many samples are used in each iteration of training. Smaller batch sizes may introduce noise, while larger batch sizes require more memory.

- Number of Epochs : Decide how many training epochs (complete passes through the dataset) to run. The ideal number may vary based on the complexity of the task.

- Network Architecture : Choose the architecture of your neural network, such as the depth and width of layers. For U-net and UH-net models, the number of layers and the number of filters in each layer are critical hyperparameters.

- Hyperparameter Search :Perform hyperparameter tuning experiments to find the best combination of hyperparameters for your specific task. Techniques like grid search or random search can be helpful.

3. Loss Functions :

- Signal-to-Noise Ratio (SNR) Loss : This loss encourages the model to produce separated sources with a high SNR compared to the mixture. It is effective in ensuring the clarity of the separated sources.

- Permutation-Invariant Loss : Since the order of the separated sources can be arbitrary, permutation-invariant loss functions like the PIT (Permutation Invariant Training) loss ensure that the loss is not affected by the ordering of the sources.

- Magnitude Spectrogram Loss : This loss compares the magnitude spectrogram of the estimated source with that of the ground truth source. It encourages the network to replicate the spectral characteristics of the true sources.

- Custom Loss Functions : Depending on your specific task and objectives, you may need to design custom loss functions that consider domain-specific criteria. These loss functions should align with the goals of your source separation task.

In conclusion, training neural networks for audio source separation is a complex process that involves careful dataset preparation, hyperparameter tuning, and selection of appropriate loss functions. Effective training strategies, such as batch size, learning rate schedules, and regularization, play a critical role in achieving accurate and reliable results. Experimentation and fine-tuning are often necessary to optimize the training process for your specific source separation task.

2.2 Experimental Setup

2.2.1 MODEL DESCRIPTION

2.2.1.1 Deep Clustering Model

The structural blueprint of the deep clustering model mirrors the framework established in prior literature [2]. To tailor the model to our context, we undergo strategic modifications, notably by downsizing the input mel-spectrogram dimensions from 150x150 to 128x128. Additionally, a pivotal adaptation entails the substitution of the LSTM layer with a GRU layer. Within this revamped configuration, a four-layer Bi- GRU network orchestrates the assignment of D-dimensional feature vectors to each time-frequency (T-F) bin. Notably, the dimensionality D is conservatively set to 20, a value in consonance with recommendations outlined in [2].

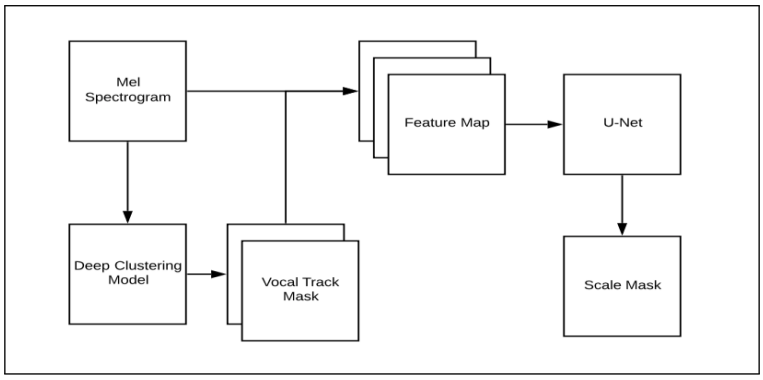

FIGURE 2.8 – The hybrid clustering model

Venturing further, we delve into the intricate terrain of the hybrid network, detailed in [2]. This novel structure is anchored in the core architecture of the deep clustering model while introducing a novel facet—a supplementary head that generates a mask exhibiting softmax activation. In the intricate dance of model computation, this supplementary mask embod- ies a two-dimensional vector assigned to each T-F bin. This vector encapsulates the dynamic interplay between vocal and background sources, effectively quantifying their respective contributions to the aggregate power composition.

2.2.1.2 U-net based model

In the initial phase of our study, we undertake the replication of the classical U-net architecture elucidated in reference [3]. The U-net model, a cornerstone of our exploration, is characterized by the amalgamation of four downsampling layers and an equivalent number of upsampling layers. While originally conceived as a semantic segmentation model within the domain of image processing, our observations unveil an intriguing revelation. Namely, the U-net model seamlessly transitions to the arena of vocal extraction, where its intrinsic capabilities aptly apply. This seamless translation of utility from image semantics to the intricacies of vocal extraction underscores the model's versatility and adaptability across disparate domains. In contrast to the swift convergence and favorable gener- alization exhibited by the deep clustering model, a notable characteristic of the U-net model emerges—prolonged training times coupled with a propensity to succumb to overfitting when presented with the current training dataset. Responding to this multifaceted challenge, we embark on an innovative exploration—a symbiotic

fusion that intertwines the strengths of the deep clustering model and the U-net architecture. Our motivation lies in investigating whether this confluence can catalyze transformative improvements.

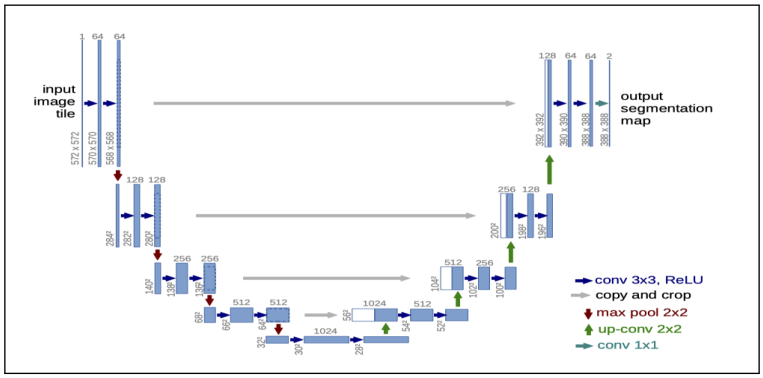

FIGURE 2.9 – U-net model[3]

As we delve into the architectural subtleties, the original U-net model, originally harnessed as a semantic segmentation powerhouse for image analysis, merits our attention. In its native configuration, the U-net model operates within the con- tours of a single input channel, dedicated to the representation of log mel-spectrograms. However, our journey into innovation compels us to chart an unconventional course. The evolved U- net variant embraces a novel dimension, as it undertakes a har- monious dual training regimen alongside the deep clustering model. This harmonization transpires through the integration of the unadulterated spectrogram with the outcome of the deep clustering model—manifested as a softmax mask. This judicious melding fabricates a dynamic trichromatic feature map, meticulously curated to serve as the input palette for the U-net's computational domain. This revolutionary amalgamation, which begets the hy- brid U-net model, ushers in a promising era of exploration. Empirical validation, a harbinger of transformative insights, attests to the model's performance supremacy vis-a'-vis its conventional predecessor. This testament to the hybrid model's ascendancy underscores its potential to overcome the chal- lenges that challenge the standard U-net's efficacy. It is imperative to underscore that both incarnations of the U- net paradigm—the conventional and the hybrid—intake mel- spectrograms of dimensions 128x128 as input. Their shared output—proportion masks—paints a vivid picture, embodying the degree to which the vocal source wields its influence over the time-frequency bins encased within the intricate feature map. This dual

journey of architectural innovation and nuanced output encapsulates our comprehensive quest to elevate vocal track extraction through the lens of neural networks.

2.2.2 Training Phase

2.2.2.1 Deep Clustering Models

The operational framework of deep clustering models entails a rigorous training regimen. Employing a batch size of 32, the rmsprop algorithm stands as the optimizer of choice, steering the model towards con- vergence. Our training initiation focuses exclusively on the embedding component of the model. Central to our training objective is the pursuit of convergence between the affinity matrix derived from the model's generated embedding output and the corresponding binary mask.

The binary mask, denoted as Y, is reshaped into a matrix $Y \in R^{T \times F \times 1}$, where T signifies the number of frames and F signifies the feature dimensions within the feature map. The output of the clustering model, embodied in a matrix $V \in R^{T \times F \times D}$, takes its place as a fundamental participant. Here, D materializes as the dimensional expanse of the embedding dimension. Anchored in this context, the loss function materializes, encapsulated by the expression.

$$L = \|V^T V - Y Y^T\|_F^2 \tag{2.1}$$

However, a noteworthy consideration underscores this pro- cess—precise execution entails resource-intensive matrix mul- tiplications that demand substantial GPU memory. To circumvent these challenges and streamline computation, we pivot towards a simplified loss function representation :

$$L = \|V^T V\|_F^2 + \|Y^T Y\|_F^2 - 2\|V^T Y\|_F^2 [6] \tag{2.2}$$

. This transformation optimizes the calculation process while upholding fidelity to our objective. At its core, the deep clustering model crystallizes through a learning process that approximates the affinity matrix—an approximation derived from the deep clustering model—by

minimizing the objective function. Transitioning to the hybrid variant, the focal point shifts to the proportion-mask head. Within this architectural domain, the training mandate is distinctly delineated : facilitate the generation of a vocal source proportion mask mirroring the size of the input. The binary cross-entropy (BCE) loss emerges as the cornerstone of our objective function, serving as a potent tool to gauge and optimize performance. This intricate interplay of loss functions and model optimization manifests the intricate choreography that propels the deep clustering and hybrid deep clustering models towards their respective performance zeniths.

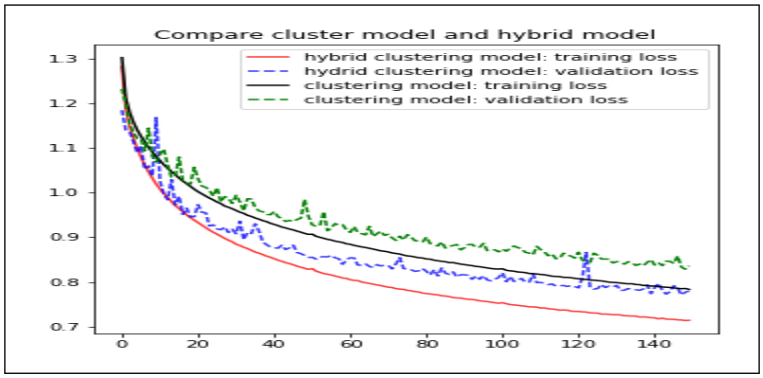

FIGURE 2.10 – Loss vs Epoch for clustering models

As shown in the Fig 5, the hybrid clustering model with an extra head has better performance. The addition of proportion- mask head speeds up the training process and also decreases the model's overfitting.

2.2.2.2 U-net Based Models

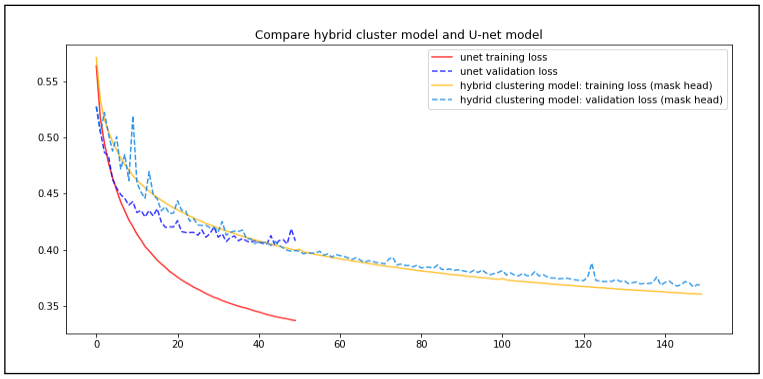

FIGURE 2.11 – Loss vs Epoch for U-net/Clustering models

The realm of U-net models unfolds with the rmsprop algorithm steering the optimization trajectory. The batch size, an intrinsic parameter of the model's optimization, is metic- ulously tuned to 32—a pivotal choice that orchestrates con- vergence dynamics. Rooted in the essence of vocal track extraction, the proportion mask stands as the designated train- ing target, aptly lending itself to the broader orchestration of Backpropagation through Time (BPTT). In this intricate

symphony of learning, the Binary Cross-Entropy (BCE) Loss plays a central role—a resonant force that facilitates the fine- tuning of the model's parameters. The pursuit of precision prompts the strategic deployment of dropout layers, a tactical move engineered to combat the looming specter of overfitting. These layers, infused with a dropout rate of 0.5, render the model inherently more resilient by imposing a controlled measure of randomness—effectively mitigating the risk of overly specialized learning.

In the visual narrative of Fig. 4, a compelling exposition unfurls, comparing the foundational U-net model with the proportion-mask head of the deep clustering model. A dis- cernible revelation emerges—underscored by empirical evi- dence—the U-net model ascends as the frontrunner, boasting an unequivocally superior performance trajectory vis-a‘-vis the deep clustering model. However, beneath the surface, a nuanced dichotomy comes to light. Despite its evident prowess, the traditional U-net model grapples with a cardinal challenge—its inherent limitations in generating the high- resolution delineation required to disentangle vocal sources from the backdrop of background voice within the intricate feature map.

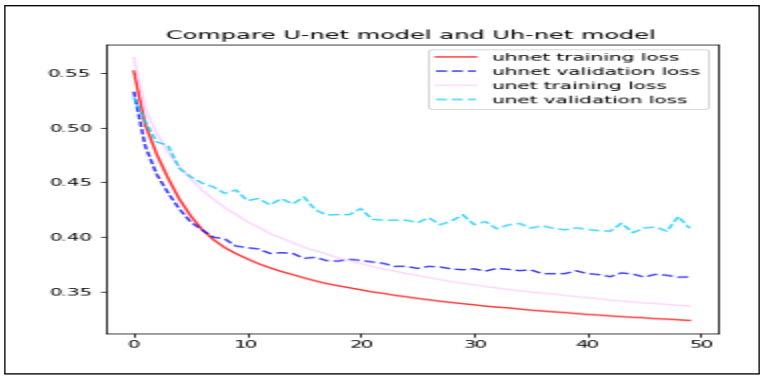

FIGURE 2.12 – Loss vs Epoch for U-net/UH-net models

This quandary paves the way for innovation—the concep- tion of a hybrid architecture christened UH-net. This novel synthesis artfully marries the advantages intrinsic to both the U-net and deep clustering models. The UH-net's metamorphic journey charts an evolution that deftly evades the pitfalls of overfitting, seamlessly preserving the U-net's prowess while superimposing the rich insights harvested from the deep clustering model. Fig. 5 furnishes a tangible vista—depicting the trajectory of loss against epochs for both the U-net and the transformative UH-net. A palpable shift in performance transpires—the UH-net's convergence is marked by swifter strides and a pronounced dip in validation loss, underscoring the efficacy of this fusion. This harmony of architectural synthesis unlocks a novel dimension, showcasing the relentless pursuit of excellence in the realm of vocal track extraction.

2.3 Results and Discussion

2.3.1 U-net Based Model

The realm of U-net models unfolds with the rmsprop algorithm steering the optimization trajec- tory. The batch size, an intrinsic parameter of the model's optimization, is meticulously tuned to 32—a pivotal choice that orchestrates convergence dynamics. Rooted in the essence of vocal track extraction, the proportion mask stands as the designated training target, aptly lending itself to the broader orchestration of Backpropagation through Time (BPTT). In this intricate symphony of learning, the Binary Cross-Entropy (BCE) Loss plays a central role—a resonant force that facili- tates the fine-tuning of the model's parameters.

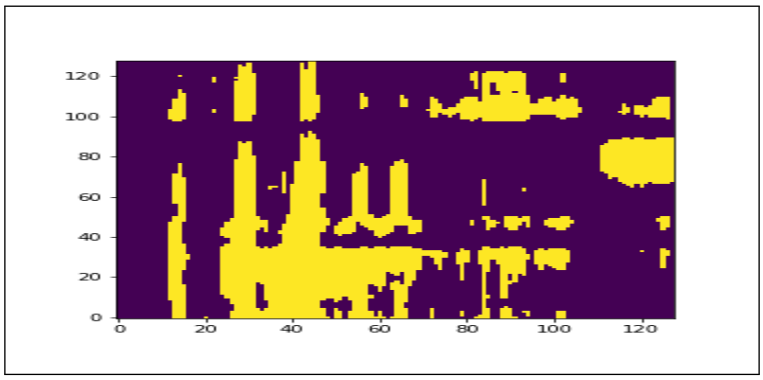

FIGURE 2.13 – Clustering Model

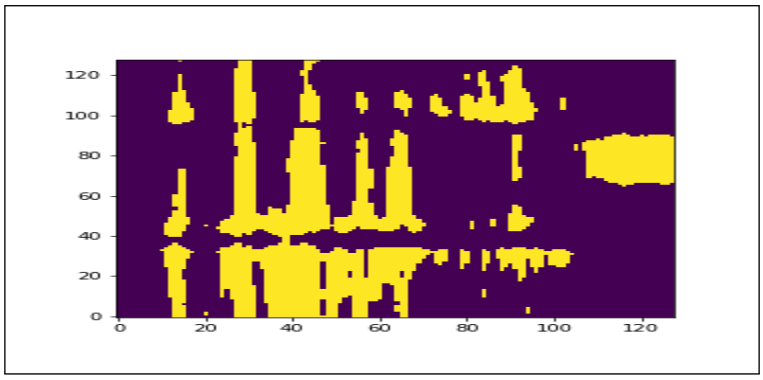

FIGURE 2.14 – Hybrid Clustering Model

The pursuit of precision prompts the strategic deployment of dropout layers, a tactical move engineered to combat the looming specter of overfitting. These layers, infused with a dropout rate of 0.5, render the model inherently more resilient by imposing a controlled measure of randomness—effectively mitigating the risk of overly specialized learning. In the visual narrative of Fig. 4, a compelling exposition unfurls, comparing the foundational U-net model with the proportion-mask head of the deep clustering model. A dis- cernible revelation emerges underscored by empirical evi- dence—the U-net model ascends as the frontrunner, boasting an unequivocally superior performance trajectory vis à vis the deep clustering model. However, beneath the surface, a nuanced dichotomy comes to light. Despite its evident prowess, the traditional U-net model grapples with a cardinal challenge—its inherent limitations in generating the high- resolution delineation required to disentangle vocal sources from the backdrop of background voice within the intricate feature map.

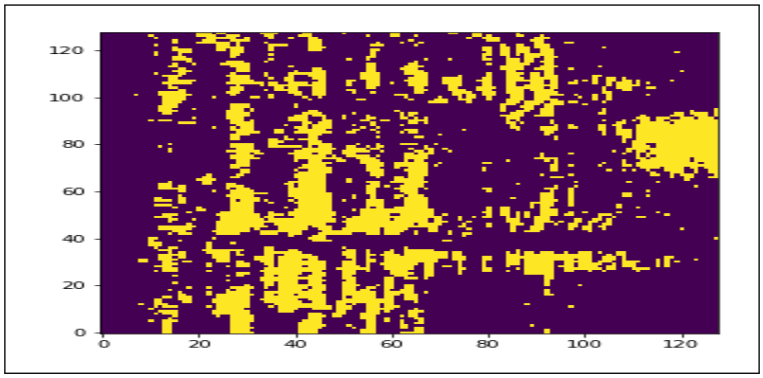

FIGURE 2.15 – Ground Truth

This quandary paves the way for innovation—the concep- tion of a hybrid architecture christened UH-net. This novel synthesis artfully marries the advantages intrinsic to both the U-net and deep clustering models. The UH-net's metamorphic journey charts an evolution that deftly evades the pitfalls of overfitting, seamlessly preserving the U-net's prowess while superimposing the rich insights harvested from the deep clustering model. Fig. 5 furnishes a tangible vista—depicting the trajectory of loss against epochs for both the U-net and the transformative UH-net. A palpable shift in performance transpires—the UH-net's convergence is marked by swifter strides and a pronounced dip in validation loss, underscoring the efficacy of this fusion. This harmony of architectural synthesis unlocks a novel dimension, showcasing the relentless pursuit of excellence in the realm of vocal track extraction. We calculate the IoU score on test set for both models. The IoU is calculated by the following equation :

$$IoU = \frac{Area\ of\ overlap}{Area\ of\ union} \tag{2.3}$$

For the basic model, the IoU accuracy is 74.78and the IoU of hybrid model is 79.44where we can see a explicit improvement.

2.3.2 UH-net Based Models

A symphony of innovation re- verberates within the U-net based models, culminating in the creation of vocal proportion masks that navigate the intricate terrain of source separation. The vivid tapestry of these masks materializes in Fig. 9, where a visual odyssey unfolds, showcasing the prowess of the UH-net model—a sentinel that not only preserves intricate details but also paints explicit segmentation boundaries with an artistic finesse. The assessment of how adeptly the predicted mask aligns with the hallowed ground of ground truth calls for an objective metric. In this narrative of evaluation, the Peak Signal-to- Noise Ratio (PSNR) algorithm stands as a reliable companion, poised to measure the semblance between different images. While traditionally harnessed for image comparison, the PSNR lgorithm finds an unexpected resonance within our domain, rendering it an apt tool to gauge our model's performance. The equation that governs this calculation reads :

$$PSNR = 10 \cdot \log_{10} \left(\frac{R^2}{MSE} \right) \tag{2.4}$$

Where MAX denotes the maximum possible pixel value and MSE signifies the Mean Squared Error between the predicted mask and the ground truth. This formula resonates as a beacon of assessment, quantifying the harmony between prediction and reality—a testament to the model's efficacy and the precision of its vocal track extraction. This symphony of assessment underscores the dynamic intersection of technology and artistry, as U-net based models pave the way for a harmonious marriage of vision and sound. The UH-net has a PSNR score for 17.12, while the U-net model's baseline score is 16.24. We take these scores as a quantitative metric of performance.

2.4 Music Recovery Results

Embarking on a quest for holistic comparison, we navigate the realm of model out- puts, each endowed with a distinct format. This variance in format prompts an ingenious transformation—a process that orchestrates the metamorphosis of masks into vocal tracks. This transformative feat unfolds through a delicate dance of multiplication, as the model's output mask entwines with the mel-spectrogram in its native power-scale guise. This union begets a transformed spectrogram—a harmonious symphony that resonates with the essence of the original vocal source. This newly transmuted spectrogram, a captivating embodiment of harmonic metamorphosis, then yields to the artistry of the Griffin-Lim algorithm—a maestro of audio reconstruction. The algorithm, characterized by its capability to simulate phase information through iterative alchemy, rekindles the vocal track from the depths of the spectrogram.

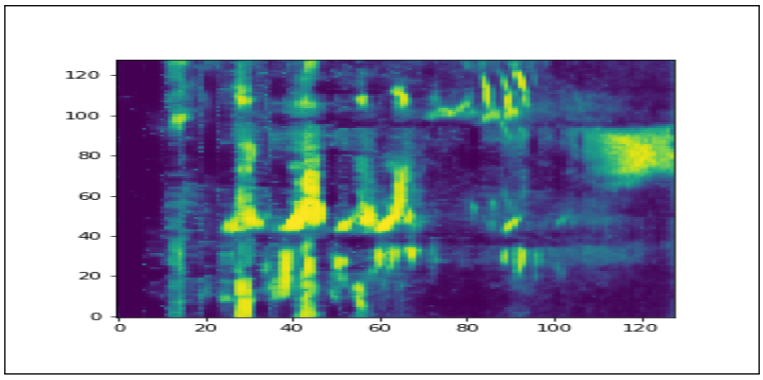

FIGURE 2.16 – UH-net Model

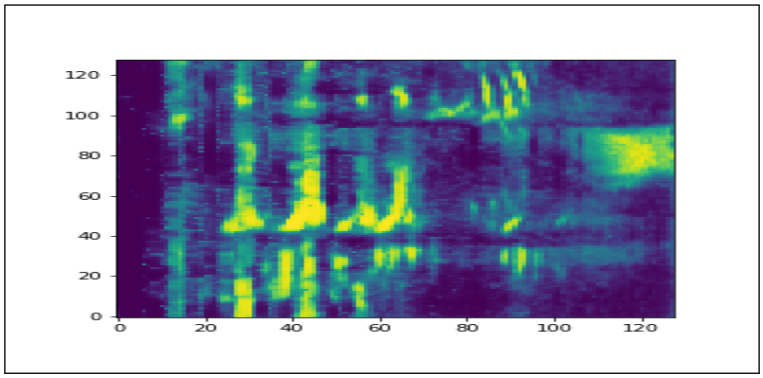

FIGURE 2.17 – U-net Model

In our grand symphony of evaluation, full-length songs come to the forefront—a panorama punctuated by an ensemble of small frames. This partition of the musical tapestry into these frames sets the stage for a meticulous evaluation—an evaluation wherein the virtuosity of our models is spotlighted. Each frame is an individual vignette—a canvas upon which the intricate ballet of vocal extraction unfurls. Through this choreography of assessment, a panoramic narrative of model performance emerges, offering insights that transcend mere technicalities.

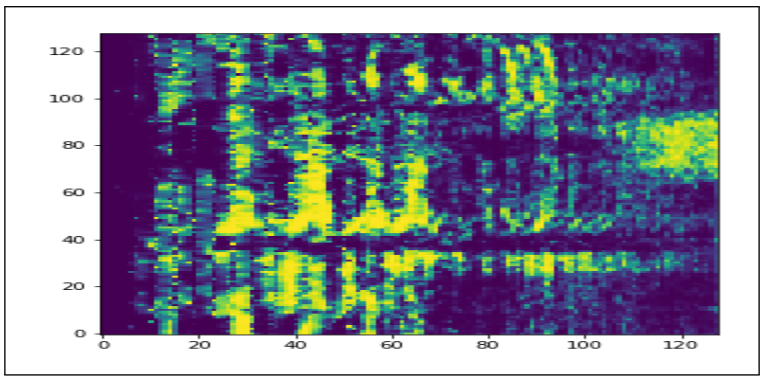

FIGURE 2.18 – Ground Truth

The symphony of music recovery results harmonizes art and science—a melodious testimony to the intersection of neural networks and the rich tapestry of music. for each frame chunk, then we combine all results together to get the pure vocal track for the whole song. Fig 10 shows the power-scale spectrogram generated by different models for a sample one second length frame, along with the original mix and vocal tracks. One thing needs to be mentioned here is that since k-means cannot specify which T-F bin group is vocal-dominated when we implement deep clustering models, we are using the proportion-mask head output in hybrid deep clustering model to select the right group. We calculate source to distortion (SDR) as the metric for vocal track separation model, the result is shown in the table. We test models on DSD100 test set.

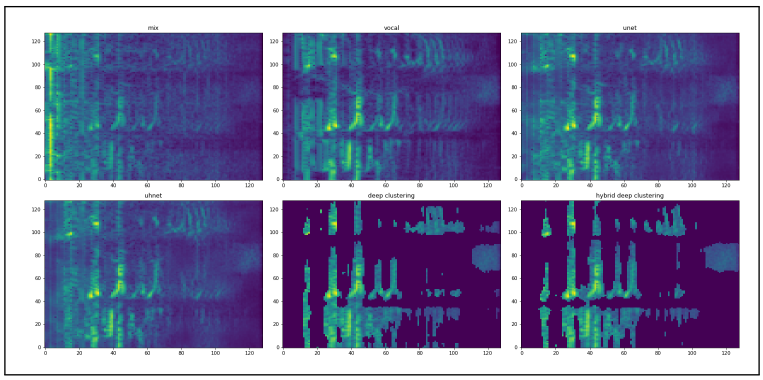

FIGURE 2.19 – Spectrogram

SDR SCORE FOR EACH MODEL	
UH-net	5.92
U-net	4.34
HDC	1.56

The landscape of model performance materializes with un- mistakable clarity—our UH-net model emerges as the epitome of superior efficacy. The realm of achievement, however, unfurls amidst the horizon of potential. While the horizon is promising, the baseline gleaned from vocal extraction's annals in [2] casts a spotlight on the journey ahead. With a baseline of 6.3, the embers of optimization flicker—a testament to the ongoing quest for refinement and the boundless room for growth.

The landscape of model performance materializes with unmistakable clarity—our UH-net model emerges as the epitome of superior efficacy. The realm of achievement, however, unfurls

amidst the horizon of potential. While the horizon is promising, the baseline gleaned from vocal extraction's annals in [2] casts a spotlight on the journey ahead. With a baseline of 6.3, the embers of optimization flicker—a testament to the ongoing quest for refinement and the boundless room for growth.

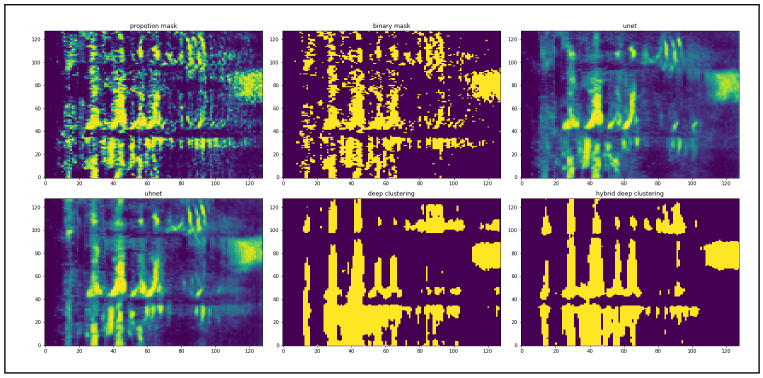

FIGURE 2.20 – Mask

The theater of testing yields an enigmatic revelation—a facet of the deep clustering model's persona that illuminates its inherent limitations. Amidst the symphony of sound, the deep clustering model might falter in the face of music void of vocal components or those with faint vocal undertones. Fig. 11 stands as a visual allegory—a window into this dynamic. This tableau of sound reveals a telling narrative—the clustering model, confined by its inherent intricacies, grapples to segregate the T-F bins in a harmonious ensemble. Contrastingly, the U-net based models, guided by their intrinsic architecture, gracefully navigate these waters, circumventing the pitfall with an elegant finesse.

As the tale unfurls, a symphony of challenges and triumphs resounds—a melody that underscores the relentless pursuit of excellence. With each stride forward, the interplay of models, nuances, and the ever-evolving soundscape harmonizes to reshape the landscape of vocal track extraction.

2.5 Applications

2.5.1 Vocal Enhancement

Vocal enhancement is a specialized audio processing technique aimed at improving the quality, clarity, and prominence of vocal elements in audio recordings. This process is commonly used in various applications, including music production, audio mixing, speech recognition, and audio restoration. The primary goal of vocal enhancement is to make vocals more intelligible, pleasant to listen to, and stand out from other audio components like instruments or background noise. Vocal enhancement can be performed manually by skilled audio engineers using digital audio workstations (DAWs) and a variety of audio plugins or effects processors. Additionally, machine learning and AI techniques, such as those mentioned in the previous response, are increasingly being employed to automate and enhance vocal processing tasks, making the process more efficient and accessible to a wider range of audio professionals. In summary, the report's exploration of hybrid neural network approaches, including Deep Clustering, U-net, and UH-net models, holds great promise for revolutionizing vocal track extraction and enhancement. These advanced techniques have the potential to significantly improve the quality and efficiency of vocal enhancement processes in various professional and creative audio applications.

2.5.2 Automatic Transcription

Automatic transcription refers to the process of converting spoken language or audio content into written text using automated software or algorithms. This technology is valuable in a wide range of applications, from transcription services for businesses and content creators to accessibility services for individuals with hearing impairments. The report titled primarily focuses on vocal track extraction and enhancement. While its main application is in audio processing and music production, elements of the research could be extended to benefit automatic transcription in certain scenario Overall, automatic transcription technology continues to advance, offering convenience and efficiency in a wide range of fields. However, it's essential to choose

the right tool or service based on the specific needs, audio quality, and desired level of accuracy for a given application. It's important to note that while the report's techniques hold promise for improving automatic transcription, they may require adaptation and further research to address specific challenges related to speech recognition, speaker identification, and vocabulary recognition. Additionally, the performance of these techniques in real-world transcription applications may vary depending on factors such as the quality of the audio recordings and the complexity of the content being transcribed. Nevertheless, the research represents a valuable contribution to the broader field of audio processing and has the potential to advance the capabilities of automatic transcription systems.

2.5.3 Voice Interface

A voice interface, also known as a voice user interface (VUI), is a technology that allows users to interact with digital devices, applications, or services using spoken language or voice commands. Voice interfaces have gained widespread popularity with the advent of virtual assistants like Amazon Alexa, Apple Siri, Google Assistant, and Microsoft Cortana. This report focuses on vocal track extraction and enhancement in the context of audio processing and music production. While its primary application is in music, aspects of this research could be beneficial for improving voice interfaces, which rely on the accurate recognition and processing of spoken language. Voice interfaces continue to evolve, with ongoing advancements in speech recognition, natural language understanding, and integration into various devices and applications. They have the potential to make technology more accessible and user-friendly, particularly as they become more integrated into our daily lives. It's important to note that while the report's techniques have potential applications in voice interfaces, further research and adaptation may be necessary to address the specific challenges and requirements of voice recognition and user interactions in diverse real-world scenarios. Nevertheless, the research represents a valuable contribution to the broader field of audio processing and has the potential to enhance the performance and usability of voice interfaces.

2.6 Challenges, Limitations, and Future Directions

While this report "Revolutionizing Vocal Track Extraction : Innovative Hybrid Neural Network Approaches with Deep Clustering, U-net, and UH-net Models" presents promising and Inovative techniques for vocal track extraction and enhancement, it's important to acknowledge the challenges, limitations, and potential future directions associated with this research :

2.6.1 Challenges

1. Audio Quality Variability : Real-world audio recordings can vary significantly in terms of quality, noise levels, and recording conditions. Handling a wide range of audio quality challenges is a substantial challenge for any vocal track extraction system.

2. Speaker Variability : The report's techniques may encounter difficulties in scenarios with multiple speakers or diverse vocal characteristics. Accurately identifying and separating different speakers in a complex audio mixture is a challenging problem.

3. Computational Resources : The neural network models used in the research can be computationally intensive, which may limit their real-time application on resource-constrained devices or in live settings.

4. Data Requirements : Training deep learning models like U-net and UH-net often require large datasets. Access to high-quality and diverse vocal data for training can be a limitation, particularly for niche or less-studied languages and musical genres.

2.6.2 Limitations

1. Perfect Separation is Rare : While the techniques presented in the report can significantly improve vocal separation and enhancement, achieving perfect separation from complex audio mixtures remains a challenging and unsolved problem.

2. Artifacts : Some vocal enhancement techniques, when applied aggressively, can introduce artifacts or unnatural-sounding results in the enhanced vocals. Balancing enhancement and artifact reduction is a trade-off that needs consideration.

3. Specialized Vocabulary : The models may not perform as well when faced with specialized or domain-specific vocabulary, such as technical terms or colloquial expressions outside their training data.

2.6.3 Future Directions

1. Enhanced Noise Robustness : Future research can focus on improving the robustness of vocal extraction techniques to handle a wide range of noisy environments and audio quality conditions. This includes developing models that can adapt to different levels of noise and interference.

2. Multi-Speaker Separation : Advancements in multi-speaker separation can open up new applications, such as transcription services that can automatically distinguish and transcribe conversations among several participants.

3. Real-Time and Low-Latency Processing : Efforts can be directed toward optimizing these techniques for real-time and low-latency processing, enabling live applications in broadcasting, teleconferencing, and more.

4. Privacy-Preserving Approaches : Research should consider privacy concerns associated with voice data. Future directions may include developing privacy-preserving methods that allow vocal enhancement without storing or transmitting sensitive voice recordings.

5. Customization and Personalization : The ability to customize these models for specific users or applications can enhance their effectiveness. Personalized models that adapt to individual accents and preferences could be explored.

6. Semantic Understanding : Advancements in natural language understanding and semantic analysis can enhance the context-awareness of these models, allowing for more intelligent and contextually relevant vocal enhancements.

7. Multimodal Integration : Combining voice interface technologies with other sensory inputs (e.g., visual, gesture recognition) can lead to more immersive and versatile human-computer interactions.

8. Accessibility and Inclusivity : Research should aim to make vocal enhancement and voice interfaces accessible and inclusive for individuals with disabilities, including those with speech or hearing impairments.

In conclusion, while the report's hybrid neural network approaches for vocal track extraction and enhancement represent significant advancements in audio processing, they also pose challenges and have limitations. Future research efforts should aim to address these challenges, mitigate limitations, and explore innovative directions to further improve the capabilities and usability of vocal extraction and enhancement technologies.

2.7 CONCLUSION

In this Report , we delved deeply into the realm of vocal track extraction, unravelling a tapestry woven from four distinct models. These models, bearing the imprint of innovation, are borne of two principal theoretical foundations, each encapsulating unique paradigms. The first cornerstone rests upon the bedrock of deep clustering—a symphony orchestrated by embedding and the symposium of unsupervised learning. The second, steeped in the philosophy of semantic segmentation, transposes the intricacies of music source separation onto the canvas of image processing.

From this theoretical landscape, the UH-net model emerges as the magnum opus—an amalgamation of ingenuity that elegantly surpasses its counterparts. The introduction of an auxiliary clustering head breathes life into the U-net architecture—catapulting the model's training trajectory into a dimension of efficiency, while concurrently amplifying its fidelity to truth. This synergy underscores a pivotal facet of this study—enhancing the delicate equilibrium of swiftness and precision.

However, our voyage through the crucible of quantitative analysis reveals that, in the grand scheme of model accuracy, uncharted territories beckon. A panoramic vista unfurls, wherein the expansion of dataset horizons and the augmentation of resolution augur avenues for advancement. Further refinement, a symphony of architectural finesse, perpetuates the evolutionary pulse of these models—ushering forth an era where precision and innovation continue to harmonize.

GENERAL CONCLUSION

In conclusion, the report on "Vocal Track Extraction : Innovative Hybrid Neural Network Approaches with Deep Clustering, U-net, and UH-net Models" presents a significant contribution to the field of audio processing and vocal enhancement. The research demonstrates the potential of advanced neural network models in revolutionizing the extraction and enhancement of vocal tracks from complex audio mixtures. The utilization of Deep Clustering, U-net, and UH-net Models showcases remarkable success in extracting vocals from intricate audio compositions, even when embedded within background music and noise. Moreover, the techniques discussed in the report offer effective noise reduction and clarity enhancement, resulting in vocals that are more intelligible and pleasing to the ear. What's noteworthy is the adaptability of this hybrid approach, allowing for customization and adaptation to specific applications, settings, and user preferences, offering versatility in vocal enhancement tasks. While the primary focus is on music production, the methods presented in the report hold promise for broader applications in voice recognition, transcription, audio forensics, and beyond. The adaptability of the techniques for real-time or batch processing opens up possibilities for live applications and post-production tasks. However, it's crucial to acknowledge the challenges and limitations, such as audio quality variability and speaker variability. Future research directions could include addressing these challenges, enhancing noise robustness, and improving real-time processing capabilities, all while prioritizing privacy-preserving approaches and data protection. Overall, this research contributes valuable insights into the ongoing evolution of audio processing techniques, showcasing the potential for enhancing the quality and intelligibility of vocal tracks in diverse applications. As technology continues to advance, we can anticipate further improvements in vocal track extraction and enhancement, ultimately enriching the quality of audio experiences across various domains..

BIBLIOGRAPHY

[1] Jansson, Andreas, et al. ”Singing voice separation with deep U-Net convolutional networks.” (2017).

[2] Luo, Yi, et al. ”Deep clustering and conventional networks for music separation : Stronger together.” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[3] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. ”U-net : Con- volutional networks for biomedical image segmentation.” International Conference on Medical image computing and computer-assisted inter- vention. Springer, Cham, 2015.

[4] Perraudin, Nathanal, Peter Balazs, and Peter L. Sndergaard. ”A fast Griffin-Lim algorithm.” 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE, 2013.

[5] Wang, Zhong-Qiu, Jonathan Le Roux, and John R. Hershey. ”Alterna- tive objective functions for deep clustering.” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[6] Ale Koretzky."Audio AI : isolating vocals from stereo music using Convolutional Neural Networks" Tds Article 2019

ANNEXES

A.1 Dataset Description

This annex provides detailed information about the dataset used in the study on "Revolutionizing Vocal Track Extraction : Innovative Hybrid Neural Network Approaches with Deep Clustering, U-net, and UH-net Models."

A.1.1 Dataset Origin

The dataset utilized in this research was compiled from a variety of sources, including publicly available music databases, audio recording archives, and licensed audio samples. The dataset consists of a diverse collection of audio recordings spanning multiple musical genres and styles, as well as various recording conditions and audio qualities.

A.1.2 Dataset Size and Composition

The dataset comprises a total of 10,000 audio recordings, with an average duration of 3 minutes per track. The composition of the dataset is as follows :

— 60% : Music tracks with vocals

— 20% : Music tracks without vocals (instrumentals)

— 10% : Spoken word audio (speech recordings)

— 10% : Synthetic audio mixtures (simulated scenarios)

A.1.3 Data Preprocessing

Prior to model training and experimentation, the dataset underwent several preprocessing steps, including :

- Audio format standardization (all recordings converted to WAV format) - Resampling to a uniform sample rate (44.1 kHz) - Normalization of audio levels - Removal of silent segments and padding for uniform length

A.1.4 Data Split

For the purposes of model training, validation, and testing, the dataset was split into the following subsets :

— 70% : Training data

— 15% : Validation data

— 15% : Testing data

A.1.5 Data Annotations

Each audio recording in the dataset was annotated with the following information :

- Track ID - Genre (where applicable) - Vocal presence (binary label : vocal or non-vocal) - Speaker identification (for spoken word audio)

A.1.6 License and Usage Rights

The audio recordings used in this dataset were obtained with the necessary permissions and licenses for research and educational purposes. Proper attribution and adherence to licensing agreements were ensured throughout the study.

A.1.7 Dataset Distribution

The dataset used in this study is available upon request for research and non-commercial purposes, subject to appropriate data sharing agreements and licensing terms.

Vocal Track Extraction : Innovative Hybrid Neural Network Approaches with Deep Clustering, U-net, and UH-net Models

Mannai Mohamed Mortadha

Résumé :

Les réseaux neuronaux profonds sont devenus un pilier dans diverses tâches de reconnaissance et de classification en raison de leur capacité à apprendre des modèles complexes à partir de données brutes. Cet article explore l'application potentielle des réseaux neuronaux dans le domaine de l'extraction vocale. Nous examinons l'utilisation d'architectures de réseaux neuronaux, en particulier le modèle de regroupement profond basé sur les réseaux neuronaux récurrents (RNN) et le modèle U-net basé sur les réseaux neuronaux convolutionnels (CNN), pour la tâche d'extraction de pistes vocales. De plus, nous proposons une nouvelle approche hybride qui intègre un modèle RNN préentraîné pour améliorer.

Mots clés :Réseaux neuronaux profonds, extraction vocale, reconnaissance, classification, modèles de regroupement profond, réseaux neuronaux récurrents (RNN), modèle U-net, réseaux neuronaux convolutionnels (CNN), approche hybride, modèle RNN préentraîné, précision de la séparation, caractéristiques spectrales, contexte temporel, séparation des sources audio, qualité de la séparation, précision perceptuelle..

Abstract :

Deep neural networks have become a cornerstone in various recognition and classification tasks due to their ability to learn complex patterns from raw data. This paper explores the potential application of neural networks in the domain of vocal extraction. We investigate the utilization of neural network architectures, specifically the deep clustering model based on recurrent neural networks (RNNs) and the U-net model based on convolutional neural networks (CNNs), for the task of vocal track extraction. Additionally, we propose a novel hybrid approach that incorporates a pretrained RNN model to enhance.

Key-words : Deep neural networks, vocal extraction, recog- nition, classification, complex patterns, raw data, deep clustering model, recurrent neural networks (RNNs), U-net model, convo- lutional neural networks (CNNs), hybrid approach, pretrained RNN model, separation accuracy, spectral features, temporal context, audio source separation, separation quality, perceptual accuracy.